Recognizing millions of consistently unidentified spectra across hundreds of shotgun proteomics datasets

Johannes Griss[1,2], Yasset Perez-Riverol[2], Steve Lewis[2], David L Tabb[3], José A Dianes[2], Noemi del-Toro[2], Marc Rurik[4,5], Mathias Walzer[4,5], Oliver Kohlbacher[4–7], Henning Hermjakob[2,8], Rui Wang[2] & Juan Antonio Vizcaíno[2]

Mass spectrometry (MS) is the main technology used in proteomics approaches. However, on average, 75% of spectra analyzed in an MS experiment remain unidentified. We propose to use spectrum clustering at a large scale to shed light on these unidentified spectra. The Proteomics Identifications (PRIDE) Database Archive is one of the largest MS proteomics public data repositories worldwide. By clustering all tandem MS spectra publicly available in the PRIDE Archive, coming from hundreds of data sets, we were able to consistently characterize spectra into three distinct groups: (1) incorrectly identified, (2) correctly identified but below the set scoring threshold, and (3) truly unidentified. Using multiple complementary analysis approaches, we were able to identify ~20% of the consistently unidentified spectra. The complete spectrum-clustering results are available through the new version of the PRIDE Cluster resource (http://www.ebi.ac.uk/pride/cluster). This resource is intended, among other aims, to encourage and simplify further investigation into these unidentified spectra.

'Untargeted' or 'discovery' proteomics approaches have become key instruments in systems biology for unraveling a sample's underlying biological functions. The most common approach in the field is shotgun proteomics. Proteins are digested using a protease, and the resulting peptides are identified using tandem mass spectrometry (MS/MS). Following this process, most proteins present in the sample can theoretically be identified and quantified[1]. However, on average, around three-quarters of spectra measured in an MS/MS experiment remain unidentified (**Supplementary Fig. 1**). Many of these spectra appear to be of high quality and are likely to have originated from peptides[2]. Search engines that use sequence databases[3–5] rely on theoretical spectra generated from a user-defined protein sequence database and a set of protein post-translational modifications (PTMs) to interpret experimental spectra. Therefore, unidentified peptides are most likely low-signal-to-noise events, not present in the sequence database used (e.g., sequence variants), or they contain unexpected PTMs.

Several alternative approaches exist that can improve the rate of identified spectra such as *de novo* sequencing[6], sequence-tagging-based approaches[7], precursor mass-tolerant sequence database searches[2], and spectral library search engines[8]. Additionally, open modification searches rely on a sequence database or spectral library but allow mass shifts to occur[9]. All of these approaches share two disadvantages which currently prevent their general use: they are computationally expensive, and they tend to return ambiguous results where more than one resulting peptide sequence can be derived from the same spectrum at equal probabilities.

Here, we use spectrum clustering at a large scale to shed light on unidentified spectra. The likelihood that sequence variants and peptides with unexpected PTMs have been observed grows with increasing numbers of data sets coming from different origins and experimental settings. Thus, unidentified spectra can be systematically studied by exploiting the continuously growing number of MS/MS data sets available. We use our approach to identify three distinct sets of spectra: (1) spectra that are incorrectly identified, (2) spectra that are correctly identified but below the set scoring threshold, and (3) spectra that are truly unidentified.

The PRIDE Archive database[10] is one of the largest public proteomics MS data repositories worldwide and part of the ProteomeXchange Consortium of proteomics resources[11]. In 2013, we introduced a spectrum-clustering algorithm that accurately grouped all identified spectra in the PRIDE Archive at the time[12]. This allowed us to identify reliable peptide spectrum matches (PSMs) within the submitted heterogeneous data. However, the amount of data in the PRIDE Archive has grown exponentially in recent years.

Here, we report the development of a novel spectrum-clustering algorithm that is highly scalable and can cluster large amounts of unidentified spectra without incurring a high degree of false-positive matching. We show that we can accurately discriminate all three

[1]Division of Immunology, Allergy and Infectious Diseases, Department of Dermatology, Medical University of Vienna, Vienna, Austria. [2]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, UK. [3]Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, Tennessee, USA. [4]Department of Computer Science, University of Tübingen, Tübingen, Germany. [5]Center for Bioinformatics, University of Tübingen, Tübingen, Germany. [6]Quantitative Biology Center, University of Tübingen, Tübingen, Germany. [7]Max Planck Institute for Developmental Biology, Tübingen, Germany. [8]National Center for Protein Sciences, Beijing, China. Correspondence should be addressed to J.G. (johannes.griss@meduniwien.ac.at) or J.A.V. (juan@ebi.ac.uk).

RECEIVED 11 DECEMBER 2015; ACCEPTED 24 MAY 2016; PUBLISHED ONLINE 27 JUNE 2016; DOI:10.1038/NMETH.3902

subsets of spectra mentioned above. Most interestingly, we were able to recognize millions of spectra in the PRIDE Archive that are commonly observed in hundreds of proteomics experiments but consistently remain unidentified.

We show how more expensive computational methods can be used to identify the true origin of many of these spectra. Using multiple complementary analysis approaches, we were able to identify roughly 20% of the originally unidentified spectra in the PRIDE Archive. These complete data are now publicly available as part of the redeveloped PRIDE Cluster resource (http://www.ebi.ac.uk/pride/cluster). We hope that this resource encourages the development of new methodologies to unravel these currently unidentified spectra.

RESULTS

Accurately clustering the complete PRIDE Archive

We developed the new 'spectra-cluster' algorithm using the Apache Hadoop framework[13,14] to reach two main goals: (1) to increase spectrum-clustering accuracy and (2) to be scalable to handle the exponential data increase in the PRIDE Archive. To increase spectrum-clustering accuracy (defined by the proportion of incorrectly clustered spectra), we developed a new method to assess the similarity between two spectra: instead of the normalized dot product that is commonly used, we employed a probabilistic scoring approach similar to that of the spectrum library search engine Pepitome[15] (see Online Methods). To evaluate the accuracy of the algorithm, we first reanalyzed 209 human data sets from the PRIDE Archive (**Supplementary Table 1**), resulting in 10 million reliably identified spectra using SpectraST at a 1% peptide false-discovery rate (FDR) (see Online Methods). On the basis of this comprehensive test data set, we concluded that the new spectra-cluster algorithm is considerably more accurate than two previous clustering algorithms, MSCluster[16] and MaRaCluster[17], and is able to process a large and heterogeneous data set (**Fig. 1** and **Supplementary Note 1**). Additionally, it is robust when handling chimeric spectra, and it shows stable accuracy with increasing cluster size (**Supplementary Note 1**).

We then clustered all identified and unidentified spectra from all publicly available 'complete' data sets[18] in the PRIDE Archive by April 2015 (including 190 million unidentified and 66 million identified spectra, for a total of 256 million spectra). First, unidentified spectra were filtered using the *de novo* search engine PepNovo's peptide filtering function, which determines whether an MS/MS spectrum represents a peptide[6] (see Online Methods). This reduced the number of unidentified spectra by 42% (from 190 million to 111 million). The remaining unidentified and identified spectra were clustered, resulting in 28 million clusters, an approximately six-fold reduction in the initial number of spectra. This analysis took 5 d on a 30-node Hadoop cluster using 340 central processing unit (CPU) cores (see Online Methods).

Validating spectrum-clustering accuracy

The original PRIDE Cluster algorithm was developed to identify reliable PSMs in the heterogeneous data submitted to the PRIDE Archive[12]. We previously found that if at least 70% of the spectra within a cluster were identified as the same peptide, these identifications could be regarded as reliable. To validate the accuracy of the new spectra-cluster algorithm, we repeated this analysis with the data set described above, which was considerably larger than the one we had previously used[12] (**Supplementary Note 2**). We again found that if at least 70% of spectra in a cluster were identified as the same peptide sequence, these identifications could be considered reliable. This indicated that the clustering accuracy remained stable in the new algorithm despite the increased amount of data. Reliable clusters were therefore defined as clusters with at least three identified spectra where at least 70% of the spectra were identified as the same peptide (**Supplementary Note 2**).

The updated PRIDE Cluster resource

The redeveloped PRIDE Cluster resource provides full access to these spectrum-clustering results and links them with the originally submitted data in the PRIDE Archive (**Supplementary Note 3**). The clustering process will be repeated and the data updated at least once a year. Data can be accessed using a RESTful Application Programming Interface (API, **Supplementary Note 3**) and a web interface (http://www.ebi.ac.uk/pride/cluster, **Supplementary Note 4**). Additionally, spectral libraries for 16 species are available (http://www.ebi.ac.uk/pride/cluster/#/libraries). Two additional filters were used to generate these libraries from reliable clusters: (1) spectra must not be dominated by a single peak, and (2) at least 20% of the total ion current must be explained through b and y ions. The PRIDE-Cluster-derived human spectral library showed comparable accuracy to the human spectral library from the National Institute of Standards and Technology (NIST, **Supplementary Note 5**).

Additionally, the complete raw spectrum-clustering results are made available for further reanalysis. The freely available open-source clustering-file-reader Java API can be used to parse the raw result files (https://github.com/spectra-cluster/clustering-file-reader). Furthermore, consensus spectra of selected subsets (e.g., large unidentified spectrum clusters for multiple species) are available as mascot generic files (MGF) and mzML files (http://www.ebi.ac.uk/pride/cluster/#/results).

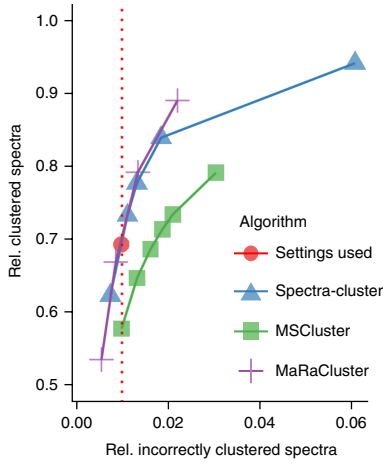

Figure 1 | Accuracy of the spectra-cluster algorithm compared with the MSCluster[16] and MaRaCluster[17] algorithms. The three algorithms were compared using a test data set built from 209 human data sets from the PRIDE Archive (see Online Methods, **Supplementary Table 1**). Clustering sensitivity (*y*-axis) was assessed based on the number of clustered spectra (shown as relative to the total number of spectra in the test data set). Clustering specificity (*x*-axis) was assessed based on the proportion of spectra that were not identified as the most common peptide in a cluster. Only clusters with at least five spectra were taken into consideration (**Supplementary Note 1**). Rel., relative.

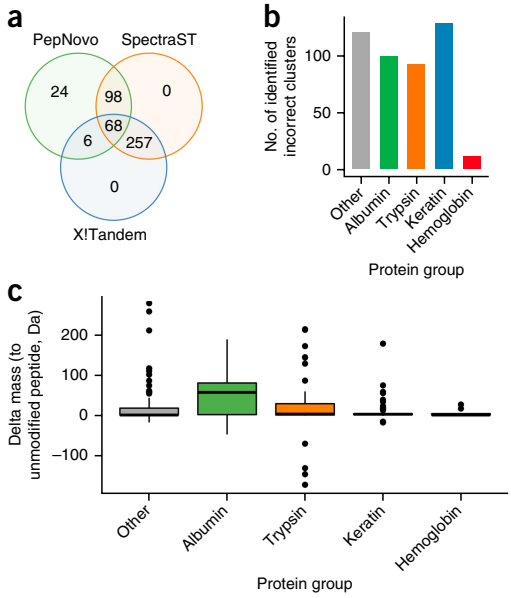

Figure 2 | Overview of the results of the analysis to highlight commonly found incorrect peptide identifications in the PRIDE Archive. (**a**) Intersection of incorrect peptide clusters in the PRIDE Archive identified using X!Tandem, SpectraST, and PepNovo. (**b**) Number of newly identified, previously incorrect clusters broken down by protein group. (**c**) Delta mass of unmodified peptides broken down by protein group (center line marks the median, edges the first and third quartile, whiskers extend to ±1.58× the interquartile ratio divided by the square root of the number of observations, and single points denote measurements outside this range).

The complete source code of the PRIDE Cluster project is available as open-source software under the permissive Apache 2.0 license at https://github.com/spectra-cluster (clustering algorithm) and https://github.com/PRIDE-Cluster (web application). Additionally, we have created a stand-alone Java application of the spectra-cluster algorithm, the spectra-cluster-cli (https://github.com/spectra-cluster/spectra-cluster-cli), which can be run on any standard computer (Windows, Mac OS, or Linux).

The presented release of the PRIDE Cluster resource (version 2015-04) includes 2.6 million reliable spectrum clusters containing 37 million identified and 9 million unidentified spectra. The relatively low number of validated originally submitted identifications (56%) is a result of the rigorous thresholds required in this highly heterogeneous data set. The vast majority of validated clusters stem from human experiments (990,853), followed by mouse (313,247), *Arabidopsis thaliana* (268,281), and rat (84,970) experiments. These correspond to 222,777; 103,834; 70,119; and 37,552 distinct peptides for human, mouse, *A. thaliana*, and rat, respectively. While the two other major repositories for MS/MS data, PeptideAtlas[19] and the Global Proteome Machine Database (GPMDB)[20], hold MS-based evidence for most of these peptides, 50,319 human, 25,101 mouse, 14,671 *A. thaliana*, and 6,873 rat peptides are only found in the PRIDE Cluster data set (**Supplementary Fig. 2**). Additionally, the list of reliable peptides identifies 870 human proteins with at least two unique peptides that have at least nine amino acids (see guidelines of the Human Proteome Project[21]) annotated without 'experimental evidence at protein level' in the human UniProtKB/SwissProt database (release 2016-03, **Supplementary Fig. 3**).

Identifying common incorrect peptide identifications

We observed 75,310 clusters (containing 2.2 million identified and 3.4 million unidentified spectra) with at least 10 identified spectra, of which less than 50% were identified as the same peptide in the original data submitted to the PRIDE Archive. Therefore, either these spectra were incorrectly clustered, or they represented peptides that are prone to be incorrectly identified.

We reprocessed the originally submitted spectra of the 3,997 large clusters containing at least 100 spectra in which at least one spectrum came from a human experiment (corresponding to 555,339 identified and 3.2 million unidentified spectra; see Online Methods, **Supplementary Fig. 4**). A spectrum cluster was accepted as identified if (1) at least two of the approaches identified the majority of spectra (more than 50% of spectra in a cluster) as the same peptide, or (2) PepNovo derived a sequence from the majority of spectra that matched a known common contaminant or proteins that are commonly found in proteomics experiments.

We were able to identify 453 clusters (11%, **Fig. 2a**). Overall, 74% of these peptides originated from keratins, trypsin, albumin, and hemoglobin (**Fig. 2b**). Albumin peptides often contained PTMs, which could explain why identifications were originally missed. Keratin, hemoglobin, and trypsin peptides, however, were mostly unmodified (**Fig. 2c**). Keratin peptides that were originally identified incorrectly were mostly found in nonhuman experiments. In these cases, the search databases that were originally used likely did not contain contaminants, which prevented the search engine from providing the correct peptide assignments for these spectra.

Inferring identifications for originally unidentified spectra through spectrum clustering

In our analysis, 9.1 million originally unidentified spectra were matched to reliably identified spectra (included in reliable clusters). These additional identifications included peptides containing biologically relevant PTMs such as phosphorylation;

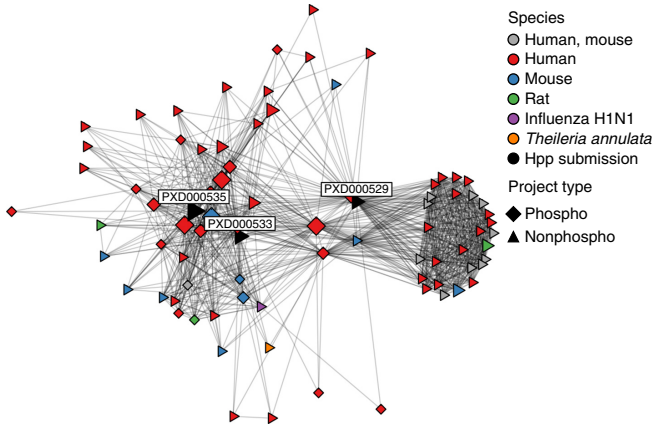

Figure 3 | Identified spectra from a diverse range of data sets, including spectra from experiments in other species, led to newly identified phosphorylated peptides in the Chromosome-Centric HPP data sets (PXD000529, PXD000533 and PXD000535). Connections between data sets are based on the shared spectra within a cluster, only taking clusters of phosphorylated peptides into consideration. Human, mouse, experiments conducted on mixed samples; phopho, studies taking phosphorylated peptides into consideration; nonphospho, studies that did not consider phosphorylated peptides.

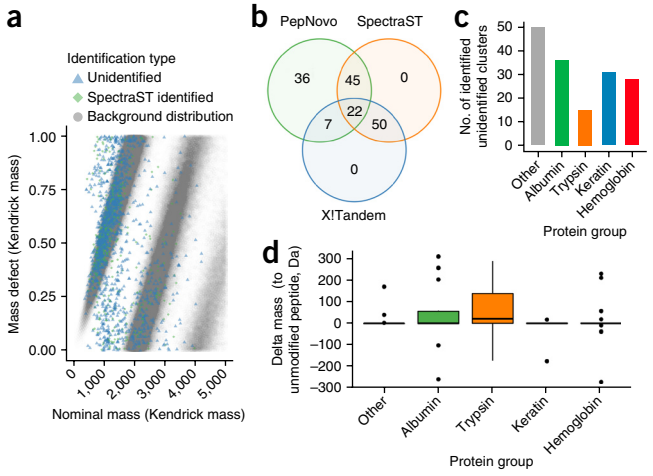

Figure 4 | Overview of the results of the analysis of clusters containing only unidentified spectra. (**a**) Mass defect analysis of unidentified spectra. (**b**) Intersection of the large unidentified human clusters identified using SpectraST, X!Tandem, and PepNovo. (**c**) Number of newly identified, previously unidentified clusters broken down by protein group. (**d**) Delta mass of unmodified peptides broken down by protein group (center line marks the median, edges the first and third quartile, whiskers extend to ±1.58× the interquartile ratio divided by the square root of the number of observations, and single points denote measurements outside this range).

49,263 reliable clusters (containing 560,000 identified and 130,000 unidentified spectra) contained phosphorylated peptides. These clusters originated from 81 phosphoproteomics studies (where enrichment for phosphopeptides was performed) and 145 nonphosphorylation studies (where no enrichment was performed, **Supplementary Table 2**). To validate these clusters, we additionally analyzed the consensus spectra of those clusters containing spectra from human experiments (28,821 clusters) using SpectraST and a human phosphopeptide spectral library[22] (see Online Methods). Overall, 8,417 (34%) of the spectra were identified at a 1% peptide FDR, of which 8,089 (96%) were identified as the most common peptide sequence in the cluster.

One example of a study we analyzed is a study included in the chromosome-centric Human Proteome Project (HPP; chromosomes 1, 8, and 20; data sets PXD000529, PXD000533, and PXD000535)[23]. Researchers analyzed the hepatocellular carcinoma cell lines Hep3B and MHCC97H and, since they did not perform any phosphoenrichment, the modification was not taken into consideration during the original analysis. However, 1,859 originally unidentified spectra were clustered with reliable identifications of phosphorylated peptides. Most of these reliably identified spectra came from five phosphorylation studies (**Fig. 3**): PXD000314 (ref. 24), a study on human lung cancer; PXD000948 (ref. 25), a study on human breast cancer; PRD000711 (ref. 26), a study on data extraction techniques; PRD000118 (ref. 27), a study on human leukocytes; and PXD000185 (ref. 28), a study on kinase substrates in leukemia cells. The 1,859 originally unidentified spectra corresponded to 290 distinct peptides (coming from 344 PSMs) and to 222 proteins (**Supplementary Table 3**), which primarily regulate translation or are involved in RNA processing and DNA repair (see Online Methods, data not shown).

A second example is a nonphosphorylation study performed by Menschaert *et al.* using mouse embryonic stem cells (PXD000124)[29]. Here, a phosphorylation study on human leukemia cells (PXD000185)[28] and on human breast cancer cells (PXD000472)[30] led to potential additional phosphopeptide identifications. Although only 82 unidentified spectra were clustered with phosphorylated peptides (**Supplementary Table 3**), this example illustrates that additional identifications are possible across different species. A more peculiar, but highly plausible, example is that a phosphorylation study on *Plasmodium falciparum* (PXD000070)[31] led to phosphopeptide identifications in a nonphosphorylation study characterizing the erythrocyte membrane human proteome (PRD000092, **Supplementary Table 3**)[32]. Since *P. falciparum* develops in human erythrocytes, it is likely that both studies contained peptides from human erythrocytes.

Analyzing clusters containing only unidentified spectra

A total of 19 million clusters (corresponding to 105 million spectra) contained only unidentified spectra. Out of these, 41,155 clusters contained more than 100 spectra (corresponding to 12.1 million spectra, 12%), indicating that these clusters may represent highly abundant molecules that have been detected (but not identified) across many different experimental settings.

A mass defect analysis[33] of all 22,344 large human clusters (taking only high-resolution spectra from Orbitrap instruments into consideration) indicated that the vast majority of these unidentified spectra originated from peptides (**Fig. 4a**). ~80% of the unidentified human spectra had a similar distribution to the background

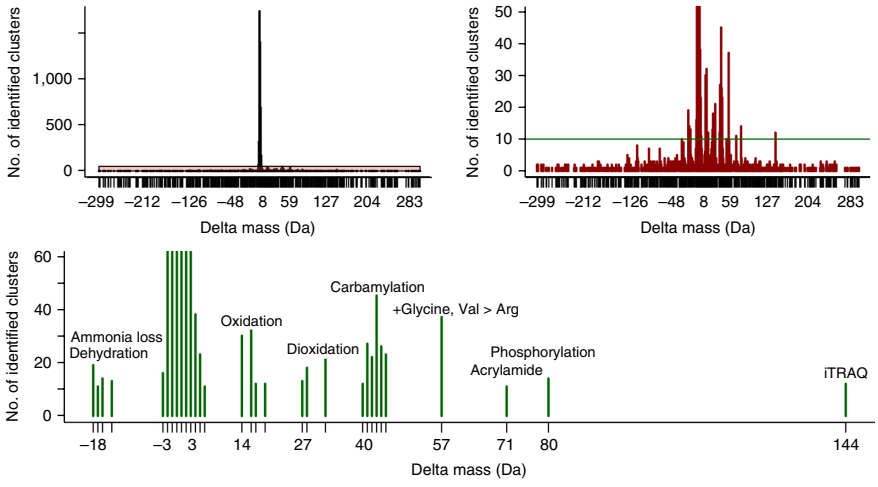

Figure 5 | Delta masses observed for the 5,560 large human unidentified clusters whose consensus spectra were identified using an open modification search. All panels show the same data but with different *y*-axes (top right and lower panel limited to 60 identified clusters). Additionally, the lower panel only shows delta masses observed for at least ten clusters. Commonly observed delta masses of known PTMs are labeled in the figure. For the complete list of observed delta masses see **Supplementary Table 4**. The origin of unlabeled delta masses (−3 to 3 Da) is unknown. +Glycine, addition of glycine; Val > Arg, substitution of valine with arginine; iTRAQ, isobaric tag for relative and absolute quantitation.

distribution created from all *in silico*–digested tryptic peptides in UniProtKB/SwissProt. The remaining 20% of spectra not included within this distribution may be explained by the fact that only unmodified, fully tryptic peptides were considered for this distribution. An additional search against a metabolite database using OpenMS[34] did not result in any reliable identifications (see Online Methods). Finally, a search using MSPLIT[35] did not reliably identify any chimeric spectra (see Online Methods, **Supplementary Fig. 5**).

We reprocessed the consensus spectra of large unidentified clusters containing spectra from human (more than 100 spectra), mouse (more than 10 spectra), and *A. thaliana* (more than 10 spectra) experiments. These amounted to 22,344 unidentified clusters for human (7 million spectra); 131,353 unidentified clusters for mouse (5 million spectra); and 8,055 unidentified clusters for *A. thaliana* (294,079 spectra). The consensus spectra were searched with SpectraST, using its open modification search function with a precursor tolerance of 500 m/z units against the corresponding NIST spectral library (for human and mouse) or, in the case of *A. thaliana*, the PRIDE Cluster spectral library (version 2015-04).

We were able to identify 5,560 human (25%), 16,439 mouse (13%), and 250 *A. thaliana* (3%) consensus spectra at a 1% peptide FDR (**Fig. 5**, **Supplementary Figs. 6** and **7**). The vast majority of newly identified peptides were detected with a mass difference (delta mass) between −2 and +4 Da, similar to previous findings[2] (**Supplementary Table 4**). Most commonly observed delta masses could be mapped to known PTMs as well as to one potential amino acid substitution (**Fig. 5**, bottom panel).

We then reprocessed the originally submitted spectra of the 1,357 human clusters with at least 1,000 spectra, 835 mouse clusters with at least 500 spectra, and 576 *A. thaliana* clusters with at least 50 spectra using the pipeline mentioned above (**Supplementary Fig. 4**). In total, 160 (12%) human, 122 (15%) mouse, and 50 (9%) *A. thaliana* clusters were identified. Most human clusters were identified as peptides corresponding to trypsin, albumin, hemoglobin, and keratin (**Fig. 4b–d**). The identifications of the mouse and *A. thaliana* clusters were more heterogeneous but could not be related to any additional protein subgroups (**Supplementary Figs. 8** and **9**). In all three species, trypsin peptides often contained PTMs such as methylation and dimethylation, which can be artificially introduced to prevent self digestion[36] (**Fig. 4c**, **Supplementary Figs. 8c** and **9c**). 242 additional human clusters had high-score PepNovo results that could not be matched to the human sequence database used.

DISCUSSION

Using our clustering method, we were able to highlight spectra consistently unidentified across thousands of experiments available in the PRIDE Archive and assign identifications to 9 million originally unidentified spectra. Additionally, we show that PTMs missed by inadequate search engine settings can be identified. The fact that many of these incorrect identifications are caused by known contaminants will be used as basis for a future service in the PRIDE Archive, which will automatically warn submitters if their data sets contain a high proportion of such potentially incorrect identifications.

However, our resource must be seen as an initial step. We are unaware of any method to aptly quantify the FDR for the 'rescued', inferred identifications shown. Nevertheless, this information can be used to reanalyze the data set of interest, taking highlighted missed PTMs or missing sequences into account. Thereby, spectrum clustering may act as an unbiased assessment of the search strategy used.

The large amount of seemingly 'good' unidentified spectra in MS/MS-based experiments is currently a core interest in the field, and many of these unidentified spectra seem to originate from modified peptides[2]. We are now able to accurately target and, in some cases, identify spectra that are observed across a multitude of experiments but remain unidentified. The majority of mass shifts that were observed could be linked to common PTMs. But the accuracy of this analysis is limited, as we only analyzed consensus spectra (see the delta mass values in **Supplementary Table 4**). As these spectra were generated from both high- and low-resolution data, the resulting mass accuracy is insufficient to accurately analyze the observed mass shifts. Still, it is intriguing that we were able to identify less than 20% of these data, and we still cannot explain a large number of observed delta masses.

The available raw clustering results represent a spectral archive of the public data in the PRIDE Archive and can be seen as a compressed data storage mechanism. This has been proposed as an ideal way to make such huge amounts of data available for reanalysis[37] with two main challenges to overcome: (1) to enable the transfer of gigabytes of data across the Internet and (2) to improve the spectrum-clustering algorithms to handle the growth in data while maintaining accuracy. The first issue has been overcome by the availability of faster file transfer protocols like Aspera (http://asperasoft.com/), routinely used by PRIDE Archive submitters. The second issue has now been tackled with the new spectra-cluster algorithm.

Our analysis only touches the tip of the iceberg. Deriving novel biological knowledge from these potential novel identifications goes far beyond the capabilities of a single research group. Creating a sensible subset of spectra to start an in-depth analysis of unidentified spectra has been very challenging. We provide these ready-to-use collections of commonly unidentified spectra as a resource to the community.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

This work was supported by the Vienna Science and Technology Fund (WWTF, grant LS11-045; grant was awarded to S.N. Wagner (Medical University of Vienna, Division of Immunology, Allergy and Infectious Diseases) and used to fund J.G.), the Wellcome Trust (grant WT101477MA to H.H. and J.A.V.), the BBSRC ('PROCESS' grant BB/K01997X/1 to H.H. and J.A.V., 'Quantitative Proteomics' grant BB/I00095X/1 to H.H.), the Deutsche Forschungsgemeinschaft (grant SFB685/B1 to O.K.), and the BMBF (grant 01ZX1301F to O.K.). We would like to acknowledge the attendees of the Midwinter Proteomics Bioinformatics Seminar 2015 at Semmering (Austria) and the Bioinformatics Hub at the HUPO conference 2015 at Vancouver (Canada), who provided valuable feedback on the data analysis. Finally, we want to acknowledge M. The and L. Käll for their support during the benchmarking of their MaRaCluster algorithm.

AUTHOR CONTRIBUTIONS

J.G. developed the clustering algorithm, ran the experiments, and performed the data analysis. D.L.T. contributed to the development of the probabilistic scoring approach. Y.P.-R. contributed to the data analysis. J.G. and R.W. developed the

Java APIs for the spectrum-clustering-analysis pipeline. S.L., R.W., and J.G. developed the Hadoop implementation. J.A.D., N.d.-T., Y.P.-R., and R.W. created the web interface and the API of the PRIDE Cluster resource. M.R., M.W., and O.K. performed the metabolite search. J.G., R.W., H.H., and J.A.V. supervised the project. J.G. and J.A.V. wrote the manuscript, with contributions from the rest of the authors.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422**, 198–207 (2003).
2. Chick, J.M. *et al.* A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. *Nat. Biotechnol.* **33**, 743–749 (2015).
3. Eng, J.K., McCormack, A.L. & Yates, J.R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5**, 976–989 (1994).
4. Perkins, D.N., Pappin, D.J., Creasy, D.M. & Cottrell, J.S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551–3567 (1999).
5. Craig, R. & Beavis, R.C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **20**, 1466–1467 (2004).
6. Frank, A. & Pevzner, P. PepNovo: *de novo* peptide sequencing via probabilistic network modeling. *Anal. Chem.* **77**, 964–973 (2005).
7. Tabb, D.L., Ma, Z.Q., Martin, D.B., Ham, A.J. & Chambers, M.C. DirecTag: accurate sequence tags from peptide MS/MS through statistical scoring. *J. Proteome Res.* **7**, 3838–3846 (2008).
8. Lam, H. *et al.* Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* **7**, 655–667 (2007).
9. Ma, C.W. & Lam, H. Hunting for unexpected post-translational modifications by spectral library searching with tier-wise scoring. *J. Proteome Res.* **13**, 2262–2271 (2014).
10. Vizcaíno, J.A. *et al.* 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res.* **44**, D447–D456 (2016).
11. Vizcaíno, J.A. *et al.* ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat. Biotechnol.* **32**, 223–226 (2014).
12. Griss, J., Foster, J.M., Hermjakob, H. & Vizcaíno, J.A. PRIDE Cluster: building a consensus of proteomics data. *Nat. Methods* **10**, 95–96 (2013).
13. Yao, Q. *et al.* Design and development of a medical big data processing system based on Hadoop. *J. Med. Syst.* **39**, 23 (2015).
14. Hodor, P., Chawla, A., Clark, A. & Neal, L. cl-dash: rapid configuration and deployment of Hadoop clusters for bioinformatics research in the cloud. *Bioinformatics* **32**, 301–303 (2016).
15. Dasari, S. *et al.* Pepitome: evaluating improved spectral library search for identification complementarity and quality assessment. *J. Proteome Res.* **11**, 1686–1695 (2012).
16. Frank, A.M. *et al.* Spectral archives: extending spectral libraries to analyze both identified and unidentified spectra. *Nat. Methods* **8**, 587–591 (2011).
17. The, M. & Kall, L. MaRaCluster: a fragment rarity metric for clustering fragment spectra in shotgun proteomics. *J. Proteome Res.* **15**, 713–720 (2016).
18. Ternent, T. *et al.* How to submit MS proteomics data to ProteomeXchange via the PRIDE database. *Proteomics* **14**, 2233–2241 (2014).
19. Desiere, F. *et al.* The PeptideAtlas project. *Nucleic Acids Res.* **34**, D655–D658 (2006).
20. Craig, R., Cortens, J.P. & Beavis, R.C. Open source system for analyzing, validating, and storing protein identification data. *J. Proteome Res.* **3**, 1234–1242 (2004).
21. Omenn, G.S. *et al.* Metrics for the Human Proteome Project 2015: progress on the human proteome and guidelines for high-confidence protein identification. *J. Proteome Res.* **14**, 3452–3460 (2015).
22. Hu, Y. & Lam, H. Expanding tandem mass spectral libraries of phosphorylated peptides: advances and applications. *J. Proteome Res.* **12**, 5971–5977 (2013).
23. Liu, Y. *et al.* Chromosome-8-coded proteome of Chinese Chromosome Proteome Data set (CCPD) 2.0 with partial immunohistochemical verifications. *J. Proteome Res.* **13**, 126–136 (2014).
24. Tsai, C.F. *et al.* Sequential phosphoproteomic enrichment through complementary metal-directed immobilized metal ion affinity chromatography. *Anal. Chem.* **86**, 685–693 (2014).
25. Ye, X. & Li, L. Macroporous reversed-phase separation of proteins combined with reversed-phase separation of phosphopeptides and tandem mass spectrometry for profiling the phosphoproteome of MDA-MB-231 cells. *Electrophoresis* **35**, 3479–3486 (2014).
26. Mancuso, F., Bunkenborg, J., Wierer, M. & Molina, H. Data extraction from proteomics raw data: an evaluation of nine tandem MS tools using a large Orbitrap data set. *J. Proteomics* **75**, 5293–5303 (2012).
27. Raijmakers, R., Kraiczek, K., de Jong, A.P., Mohammed, S. & Heck, A.J. Exploring the human leukocyte phosphoproteome using a microfluidic reversed-phase-TiO2-reversed-phase high-performance liquid chromatography phosphochip coupled to a quadrupole time-of-flight mass spectrometer. *Anal. Chem.* **82**, 824–832 (2010).
28. Casado, P. *et al.* Kinase-substrate enrichment analysis provides insights into the heterogeneity of signaling pathway activation in leukemia cells. *Sci. Signal.* **6**, rs6 (2013).
29. Menschaert, G. *et al.* Deep proteome coverage based on ribosome profiling aids mass spectrometry-based protein and peptide discovery and provides evidence of alternative translation products and near-cognate translation initiation events. *Mol. Cell. Proteomics* **12**, 1780–1790 (2013).
30. Casado, P., Bilanges, B., Rajeeve, V., Vanhaesebroeck, B. & Cutillas, P.R. Environmental stress affects the activity of metabolic and growth factor signaling networks and induces autophagy markers in MCF7 breast cancer cells. *Mol. Cell. Proteomics* **13**, 836–848 (2014).
31. Collins, M.O., Wright, J.C., Jones, M., Rayner, J.C. & Choudhary, J.S. Confident and sensitive phosphoproteomics using combinations of collision induced dissociation and electron transfer dissociation. *J. Proteomics* **103**, 1–14 (2014).
32. van Gestel, R.A. *et al.* Quantitative erythrocyte membrane proteome analysis with Blue-native/SDS PAGE. *J. Proteomics* **73**, 456–465 (2010).
33. Sleno, L. The use of mass defect in modern mass spectrometry. *J. Mass Spectrometry* **47**, 226–236 (2012).
34. Sturm, M. *et al.* OpenMS - an open-source software framework for mass spectrometry. *BMC Bioinformatics* **9**, 163 (2008).
35. Wang, J., Pérez-Santiago, J., Katz, J.E., Mallick, P. & Bandeira, N. Peptide identification from mixture tandem mass spectra. *Mol. Cell. Proteomics* **9**, 1476–1485 (2010).
36. Schittmayer, M., Fritz, K., Liesinger, L., Griss, J. & Birner-Gruenberger, R. Cleaning out the litterbox of proteomic scientists' favorite pet: optimized data analysis avoiding trypsin artifacts. *J. Proteome Res.* **15**, 1222–1229 (2016).
37. Lam, H. Spectral archives: a vision for future proteomics data repositories. *Nat. Methods* **8**, 546–548 (2011).

ONLINE METHODS

Test data set and assessment of clustering accuracy. To validate the spectrum-clustering algorithm, 209 human data sets from the PRIDE Archive were reprocessed (**Supplementary Table 1**). The submitted identified spectra were searched using SpectraST[8] (version 5.0) with a precursor tolerance of 3 *m*/*z* units, ignoring the original charge states, and enabling the calculation of *P*-values. All other settings were left at their default values. As spectral library, we used a combination of the NIST human Orbitrap (November 2014) and iontrap (May 2014) libraries and the Global Proteome Machine's (GPM) common Repository of Adventitious Proteins (cRAP) (downloaded on July 2014). The libraries were combined and decoy spectra appended using SpectraST. Roughly 10 million spectra were identified at a 1% peptide FDR.

The identified spectra were clustered using the spectra-cluster algorithm and the MSCluster[16] and MaRaCluster[17] algorithms as benchmarks. Sensitivity was assessed based on the proportion of clustered spectra (spectra clustered with at least one other spectrum). Specificity was assessed based on the proportion of spectra identified as a different peptide than the most common peptide identification reported in the cluster. The MSCluster algorithm was run using the default settings, which are optimized for heterogeneous, low-resolution data. The MaRaCluster algorithm was set to a precursor tolerance of 1,000 p.p.m. All other parameters were left at the default setting.

Probabilistic spectrum comparison. A probabilistic method was developed to assess the similarity between two spectra. The approach is based on the scoring function used in Pepitome[15] (**Supplementary Note 6**). First, precursor peaks are removed from the MS/MS spectrum and peak picking is performed, keeping the 70 highest peaks per spectrum (**Supplementary Note 7**). Next, the probability that the number of matched peaks occurred randomly is modeled using a hypergeometric distribution. The probability that the rank distribution of matched peaks occurred by chance is assessed using the Kendall's Tau correlation. For both test points only probabilities are calculated, instead of cumulative probabilities. This sacrifices mathematical accuracy for improved speed, which is essential for the clustering process of large amounts of spectra. The two probabilities are then combined using Fisher's method[38] and reported as the negative logarithm. This comparison is only performed using the peaks that explain 50% of the total ion current (of the prefiltered spectrum) or at least the 25 highest peaks. The consensus spectrum is then built using all peaks.

'MapReduce' adaptation of the MSCluster algorithm. The algorithm's logic follows the MSCluster approach developed by Frank *et al.*[16], adapted to the requirements of the MapReduce programming model using our probabilistic spectrum-comparison method instead of the normalized dot product (**Supplementary Note 7**). First, as described above, peak filtering is performed in a pure mapping job. Next, five successive rounds of clustering are performed with decreasing similarity thresholds to reach a final accuracy of 99%. Spectra are mapped into bins depending on the precursor's *m*/*z* value. In the initial round, a bin width of 0.2 *m*/*z* units is used. Subsequent rounds are repeated using a bin width of 4 *m*/*z* units. Finally, the five rounds of clustering are repeated, offsetting the bins by half of the used bin width. After this, the overlapping regions of the previous bins are processed together. In the first and second rounds, only spectra that share one of their five highest peaks are compared. In subsequent rounds, only spectra (or the corresponding clusters) that were among the 30 previously highest-scoring matches are compared.

Similarly to the MSCluster algorithm, an empirically derived cumulative distribution function is used to adapt the clustering threshold based on the number of comparisons (**Supplementary Note 8**). This prevents a decreased clustering accuracy caused by the multiple testing problem when processing very large data sets. The cumulative distribution function was derived by comparing randomly selected spectra from the test data set (>10 billion comparisons). Spectra were considered different if they were originally identified as different peptides and had a precursor mass difference of at least 4 *m*/*z* units. Thereby, the proportion of incorrectly matched spectra at given similarity scores could be estimated.

Code availability. The complete source code of the PRIDE Cluster project is available as open-source software under the permissive Apache 2.0 license at https://github.com/spectra-cluster (clustering algorithm, Hadoop implementation) and https://github.com/PRIDE-Cluster (web application). A stand-alone Java application of the spectra-cluster algorithm, the spectra-cluster-cli, is available at https://github.com/spectra-cluster/spectra-cluster-cli.

Identifying consensus spectra from reliable phosphopeptide spectral clusters. Consensus spectra of reliable human clusters representing phosphopeptides were searched using SpectraST (version 5.0)[8] against the human phospho library from PeptideAtlas[19,22] (version 2013-07-15). The precursor tolerance was set to 3.0 *m*/*z* units, the spectra's charge states were ignored, and the calculation of *P*-values was enabled and used for peptide FDR filtering at 1% FDR.

Identifying incorrectly identified and unidentified spectra. Consensus spectra or originally submitted spectra were processed using SpectraST[8] (version 5.0), X!Tandem[5] (version Sledgehammer, 2013.09.01.1), and PepNovo[6] (release 20101117).

For SpectraST, the spectral library was either the combined human spectral library used for the test data set or the cRAP spectral library alone. Decoy spectra were appended using SpectraST. All SpectraST searches were performed using a precursor mass tolerance of 500 *m*/*z* units. The open modification search option ignored the spectra's charge states and enabled the calculation of *P*-values, which were used for FDR filtering (**Supplementary Fig. 4**).

For X!Tandem, either the cRAP sequence database alone (downloaded on July 2014) or the concatenation of cRAP and UniProt's human proteome (2014-07) were used. The precursor tolerance was set to 3 *m*/*z* units, fragment tolerance to 0.4 *m*/*z* units, and the refinement mode was disabled. Carbamidomethylation was set as fixed modification and oxidation of M- and N-terminal acetylation as variable modifications. If the search was performed only against the cRAP library the following additional variable modifications were taken into consideration: formylation on S and K, deamidation on N and Q, carboxylation on K, and N-terminal methylation. All other settings were the default ones.

PepNovo was set to use the 'CID_IT_TRYP' fragmentation model. Allowed protein modifications were defined as follows: carbamidomethylation; oxidation on M, phosphorylation on S, T, and Y; acetylation on T, S, Y, and N-terminal; methylation on C, H, K, N, Q, R, and N-terminal; and formylation on K, S, and T. The ten highest-scoring solutions were taken into consideration. Additionally, the best-scoring as well as the most common solution across all spectra in one cluster were considered for the analysis.

MSPLIT[35] (version 1.0) was used to identify spectra originating from more than one peptide. Precursor tolerance was set to 3.0 Da, results were filtered at 1% peptide FDR using the application's spectrumMatchClassify.pl script, and the above-mentioned spectral library was used for the search.

The pathway over-representation analysis was performed using the PANTHER[39] Overrepresentation Test (release 20150430, PANTHER version 10.0 released 2015-05-15) with 'Homo sapiens (all genes)' as reference list and 'PANTHER GO-Slim Biological Process' as annotation data set. *P*-values were adjusted using a Bonferroni correction.

The mass defect analysis was performed on recalculated consensus spectra, taking only spectra from Orbitrap instruments into consideration. Then, the nominal and fractional masses of the clusters' precursor ions were plotted as previously described[33]. The background distribution was created based on an *in silico* tryptic digest of the UniProtKB/SwissProt database (release 2013-09). These high-resolution consensus spectra were additionally searched against MassBank using its Simple Object Access Protocol (SOAP) API (performed on September 2015).

38. Mosteller, F., Winsor, C.P. & Fisher, C.H. Questions and Answers. *Am. Stat.* **2**, 18–19 (1948).
39. Mi, H., Muruganujan, A. & Thomas, P.D. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res.* **41**, D377–D386 (2013).

doi:10.1038/nmeth.3902